KORYXA / ASSISTANTS IA

Créer et livrer son premier assistant IA

Extrait découverte

Chapitres 1 et 2 : cadrer le besoin du client et comprendre le RAG.

LAMADOKOU Yayra

Édition Cohere. Projet pédagogique fictif Horizon Numérique. Les pages extraites conservent la pagination du guide complet.

Chapitre 1 - Transformer une demande en projet livrable

Partir de la question que le visiteur se pose

« Je veux un chatbot » décrit un objet, mais ne précise pas encore son utilité. Pour cadrer le projet, demandez quel effort doit disparaître. Dans notre cas, une personne intéressée par une formation doit pouvoir connaître son tarif et ses conditions d’inscription sans parcourir toute la documentation. Le secrétariat doit aussi pouvoir retirer une information obsolète.

Choisissez quelques situations concrètes. Une première personne demande le prix total d’Analyse de Données. Une deuxième veut savoir si elle sera remboursée en annulant cinq jours avant la rentrée. Une troisième écrit seulement « Combien coûte la formation ? ». Ces trois demandes ne nécessitent pas le même comportement. La première attend un montant, la deuxième une règle et ses conditions, la troisième une précision.

Le cahier des charges doit donc décrire les comportements, et pas seulement nommer les technologies. Écrire « interface Next.js et moteur RAG » ne permet pas au client de juger si le résultat répond au besoin. Écrire « le visiteur obtient le montant total, le détail des tranches et le document source » fournit un résultat vérifiable.

Définir qui valide les informations

Le développeur organise le traitement des documents ; le responsable du centre valide leur contenu métier. Avant l’indexation, identifiez la personne qui fait foi sur les prix, les dates et les règlements. Un assistant ne résout pas une contradiction entre deux fichiers simplement parce qu’il utilise une IA. Le client doit choisir quelle version est actuelle.

Dans Horizon Numérique, le corpus comprend un catalogue, une grille de tarifs, un règlement, un calendrier et une FAQ. Les originaux sont conservés dans la base. Cette décision facilitera la consultation des sources et la sauvegarde de l’ensemble documentaire. L’administrateur est chargé des remplacements ; les visiteurs disposent seulement de fonctions de lecture et de discussion.

Précisez également les actions autorisées. Notre assistant informe, cite et oriente vers le secrétariat. Il ne confirme ni paiement, ni place, ni exception de remboursement. Ces opérations exigeraient des intégrations et des règles métier supplémentaires. Le parcours proposé correspond à une application documentaire pour une seule organisation.

Choisir des critères de réception

Un critère utile comprend une situation, un comportement attendu et une manière de le vérifier. Pour le tarif d’Analyse de Données, le montant attendu est 650 euros. Les tranches sont 220, 220 puis 210 euros. Les frais administratifs de 40 euros sont inclus. Une réponse qui ajoute ces frais au total échoue au contrôle, même si elle cite correctement la grille tarifaire.

Pour l’annulation cinq jours avant la rentrée, la réponse doit tenir compte de la règle des sept jours et de l’absence de remboursement dans ce cas. Elle peut expliquer qu’un report pour force majeure est

examiné par le secrétariat, sans le présenter comme automatique. Pour la question ambiguë sur « la formation », le bon résultat est une demande de précision.

Mesurez aussi le temps de réponse dans des conditions décrites : machine, connexion, nombre de requêtes et modèle. Une moyenne de quatre secondes dans un devis est un objectif à vérifier, pas un résultat déjà obtenu. Évitez de transformer un nombre choisi au cadrage en garantie technique sans mesure.

Réalisation et vérification

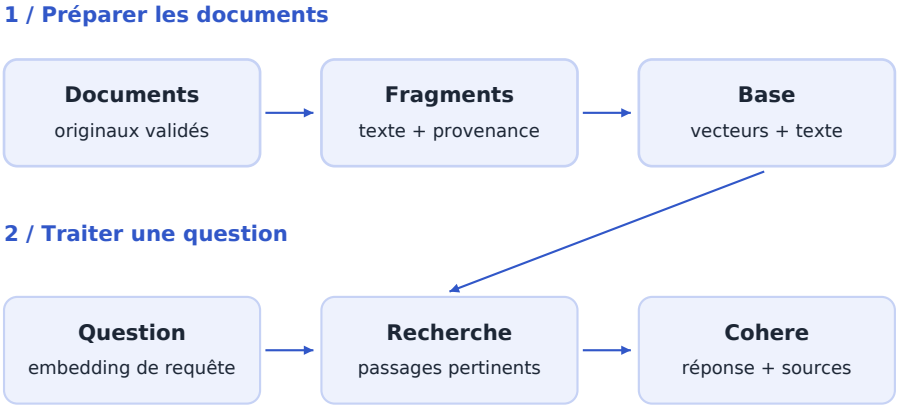
Ouvrez `modeles_clients/01_cahier_des_charges_modele.md`. Complétez l’objectif, les utilisateurs, le corpus et le responsable documentaire. Écrivez dix demandes représentatives avant d’ajouter une fonctionnalité. Puis classez chaque demande : réponse factuelle, précision nécessaire, information absente ou demande d’action.

Vérifiez que le client comprend ce classement. Si la même demande appelle deux résultats incompatibles selon les personnes interrogées, le besoin doit encore être clarifié. Une difficulté repérée à ce stade coûte moins de travail qu’une règle changée au moment de la livraison.

Exercice

Le client demande : « L’assistant doit s’occuper des inscriptions ». Reformulez cette phrase en trois comportements précis. Proposition de correction : expliquer les pièces à fournir ; afficher la date limite du premier paiement ; orienter vers le secrétariat pour confirmer la place. Ces comportements correspondent au périmètre documentaire du livre.

Chapitre 2 - Comprendre le RAG à partir d’une question



Deux circuits reliés par un corpus documenté. La recherche fournit les passages au modèle de génération.

Donner au modèle les informations nécessaires

Un modèle de langage peut produire un texte fluide sans connaître les documents actuels de votre client. Le prix d’une formation fictive ne peut pas être déduit de la manière dont d’autres centres fixent leurs tarifs. L’application doit d’abord retrouver l’information pertinente, puis la présenter au modèle avec une consigne claire.

Le RAG, pour Retrieval-Augmented Generation, organise cette association entre recherche et génération. Dans le projet, la recherche sélectionne des passages du corpus. Cohere reçoit ces passages et prépare une réponse. Le serveur contrôle ensuite le format de cette réponse et la relie aux sources conservées dans la base.

Il existe deux circuits. Le premier prépare les documents lors de leur ajout ou de leur remplacement. Le second traite une question posée par un utilisateur. Séparer ces circuits évite d’extraire et de réindexer tous les fichiers à chaque conversation.

Le circuit documentaire

Un PDF est lu page par page. Le texte est découpé en fragments assez courts pour être retrouvés et transmis au modèle. Chaque fragment garde le nom de son document et sa page. Pour le Markdown, la provenance correspond à une section. Une information de provenance doit être conservée au moment de l’extraction ; la reconstituer ensuite à partir du texte généré serait fragile.

Les fragments sont envoyés au modèle d’embeddings avec `input_type=search_document`. Un embedding est une liste de nombres qui représente des propriétés du texte pour la recherche. Il ne s’agit ni d’un résumé lisible, ni d’une preuve de vérité. Deux passages proches dans cet espace peuvent parler d’un sujet similaire sans répondre à la même question.

Les vecteurs sont enregistrés avec les fragments. Dans cette édition, les vecteurs d’Embed v4 ont 1024 dimensions. Le schéma de la base et les réponses reçues de l’API sont contrôlés pour conserver cette cohérence. Changer de modèle implique de réexaminer et de réindexer le corpus.

Le circuit d’une question

Prenons « Quel est le tarif total d’Analyse de Données ? ». L’application produit un embedding de requête avec `input_type=search_query`. Elle compare ce vecteur aux fragments et effectue aussi une recherche par mots-clés. Les résultats sont réunis et classés ; quelques passages sont transmis au modèle de génération.

Le modèle reçoit des données et une consigne : répondre à partir des passages, demander une précision si nécessaire et signaler l’absence d’information. Il doit aussi indiquer les identifiants des passages utilisés. Le serveur vérifie que ces identifiants appartiennent bien aux résultats récupérés avant d’afficher les sources.

Cette vérification rend la provenance contrôlable. Elle ne démontre pas à elle seule que chaque phrase de la réponse est correctement soutenue. Un passage contenant « 650 euros » et « 40 euros inclus » pourrait être mal interprété. C’est la raison pour laquelle l’évaluation distingue le document trouvé et la réponse formulée.

Diagnostiquer trois types d’erreurs

Si le bon prix n’est pas dans le corpus, la correction porte sur le document. Si le document est présent mais que la recherche remonte un autre parcours, la correction porte sur la récupération. Si le passage est correct mais que la réponse ajoute les frais inclus, la correction porte sur l’interprétation et les consignes.

Évitez de changer simultanément le découpage, le nombre de résultats et le prompt lorsque vous enquêtez sur une erreur. Conservez la question, les passages obtenus et la réponse. Modifiez un élément, puis rejouez le même cas. Vous pourrez expliquer ce qui a effectivement amélioré le résultat.

Exercice

Une réponse annonce 720 euros pour Analyse de Données. Le corpus indique pourtant 650 euros pour ce parcours et 720 euros pour Automatisation & IA. Avant de réécrire le prompt, consultez les fragments récupérés. Si seul le tarif d’Automatisation & IA est présent, le premier problème est la recherche. Si les deux prix sont présents, examinez la formulation de la question et le contexte transmis.

CONTINUER LE PARCOURS

Du besoin client à la livraison.

La suite du guide accompagne la préparation des documents, la recherche, la génération avec Cohere, l’API et l’interface.

Vous aborderez aussi

- La vérification des réponses et de leurs sources.
- Le remplacement des documents.
- Le déploiement, la sauvegarde et la restauration.
- Le budget et la remise du projet au client.

Le guide complet est accompagné du code, des documents d’exercice, des questions de référence et des modèles de livraison.

KORYXA

Apprendre à construire. Savoir livrer.